

WHAT IS CLAIMED IS:

1        1. A method for matching patterns in a string of  
2 symbols comprising:

3            identifying a first pattern of symbols to be matched,  
4 wherein the first pattern contains a prefix pattern, a value  
5 pattern and a suffix pattern;

6            identifying candidate matches for the first pattern in  
7 the string, wherein each candidate match for the first pattern  
8 includes a candidate match for the prefix pattern, a candidate  
9 match for the suffix pattern and a candidate match for the  
10 value pattern;

11            determining a cost associated with each of the candidate  
12 matches for the first pattern, wherein the cost associated  
13 with each of the candidate matches for the pattern includes a  
14 cost associated with the corresponding candidate match for the  
15 prefix pattern, a cost associated with the candidate match for  
16 the suffix pattern and a cost associated with the candidate  
17 match for the value pattern; and

18            selecting one or more candidate matches for the pattern  
19 that meet a cost selection criterion.

1        2. The method of claim 1 wherein determining a cost  
2 associated with each of the candidate matches comprises  
3 calculating a corresponding edit distance.

1        3. The method of claim 1 wherein identifying the first  
2 pattern comprises providing a single example string wherein  
3 the first pattern is selected from the example string.

1        4. The method of claim 1 further comprising examining  
2 the string to identify spans of interest, wherein each of the  
3 spans of interest meets a specified filtering criterion.

1        5. The method of claim 4 wherein the specified  
2 filtering criterion comprises the inclusion of a keyword.

1       6. The method of claim 1 wherein selecting one or more  
2 candidate matches for the pattern that meet a cost selection  
3 criterion comprises selecting one or more candidate matches  
4 that have corresponding costs which fall below a selected  
5 threshold.

1       7. The method of claim 1 wherein selecting one or more  
2 candidate matches for the pattern that meet a cost selection  
3 criterion comprises selecting a predetermined number of  
4 candidate matches that have the lowest corresponding costs.

1       8. The method of claim 1 wherein selecting one or more  
2 candidate matches for the pattern that meet a cost selection  
3 criterion comprises selecting a candidate match that has a  
4 lowest cost and selecting additional candidate matches that  
5 have corresponding costs which are within a predetermined  
6 tolerance of the lowest cost.

1       9. The method of claim 1 further comprising adjusting  
2 the cost selection criterion and selecting one or more  
3 candidate matches for the pattern that meet the adjusted cost  
4 selection criterion.

1       10. The method of claim 1 wherein the cost associated  
2 with the corresponding candidate match for the prefix pattern,  
3 and the cost associated with the candidate match for the  
4 suffix pattern are more heavily weighted than the cost  
5 associated with the candidate match for the value pattern.

1       11. The method of claim 1 wherein the cost associated  
2 with each of the candidate matches for the first pattern is  
3 determined by adding the cost associated with the  
4 corresponding candidate match for the prefix pattern, the cost  
5 associated with the candidate match for the suffix pattern and  
6 the cost associated with the candidate match for the value  
7 pattern.

1       12. The method of claim 1 wherein identifying each  
2 candidate match for the first pattern comprises identifying  
3 the candidate match for the prefix pattern, wherein the  
4 candidate match for the prefix pattern defines a first end of  
5 a value window, then identifying a corresponding candidate  
6 match for the suffix pattern, wherein the candidate match for  
7 the suffix pattern defines a corresponding second end of the  
8 value window, wherein the candidate match for the value  
9 pattern comprises the symbols within the value window.

1       13. The method of claim 1 further comprising filtering  
2 the candidate match for the value pattern using a keyword.

1       14. The method of claim 1 further comprising filtering  
2 the candidate match for the value pattern using a regular  
3 expression.

1       15. The method of claim 1 wherein identifying candidate  
2 matches for the prefix pattern comprises constructing an edit  
3 distance matrix for the prefix pattern and identifying one or  
4 more candidate matches for the prefix pattern, constructing an  
5 edit distance matrix for the suffix pattern and identifying  
6 one or more candidate matches for the suffix pattern, and  
7 identifying a candidate match for the value pattern between  
8 each pair of candidate prefix matches and candidate suffix  
9 matches.

1       16. A computer readable medium containing instructions  
2 which are configured to implement the method comprising:  
3       identifying a first pattern of symbols to be matched,  
4 wherein the first pattern contains a prefix pattern, a value  
5 pattern and a suffix pattern;  
6       identifying candidate matches for the first pattern in  
7 the string, wherein each candidate match for the first pattern  
8 includes a candidate match for the prefix pattern, a candidate

9 match for the suffix pattern and a candidate match for the  
10 value pattern;

11 determining a cost associated with each of the candidate  
12 matches for the first pattern, wherein the cost associated  
13 with each of the candidate matches for the pattern includes a  
14 cost associated with the corresponding candidate match for the  
15 prefix pattern, a cost associated with the candidate match for  
16 the suffix pattern and a cost associated with the candidate  
17 match for the value pattern; and

18 selecting one or more candidate matches for the pattern  
19 that meet a cost selection criterion.

20  
21 17. The computer readable medium of claim 16 wherein  
22 determining a cost associated with each of the candidate  
23 matches comprises calculating a corresponding edit distance.

24  
25 18. The computer readable medium of claim 16 wherein  
26 identifying the first pattern comprises providing a single  
27 example string wherein the first pattern is selected from the  
28 example string.

1 19. The computer readable medium of claim 16 further  
2 comprising examining the string to identify spans of interest,  
3 wherein each of the spans of interest meets a specified  
4 filtering criterion.

1 20. The computer readable medium of claim 15 wherein the  
2 specified filtering criterion comprises the inclusion of a  
3 keyword.

1 21. The computer readable medium of claim 16 wherein  
2 selecting one or more candidate matches for the pattern that  
3 meet a cost selection criterion comprises selecting one or  
4 more candidate matches that have corresponding costs which  
5 fall below a selected threshold.

1        22. The computer readable medium of claim 16 wherein  
2 selecting one or more candidate matches for the pattern that  
3 meet a cost selection criterion comprises selecting a  
4 predetermined number of candidate matches that have the lowest  
5 corresponding costs.

1        23. The computer readable medium of claim 16 wherein  
2 selecting one or more candidate matches for the pattern that  
3 meet a cost selection criterion comprises selecting a  
4 candidate match that has a lowest cost and selecting  
5 additional candidate matches that have corresponding costs  
6 which are within a predetermined tolerance of the lowest cost.

1        24. The computer readable medium of claim 16 further  
2 comprising adjusting the cost selection criterion and  
3 selecting one or more candidate matches for the pattern that  
4 meet the adjusted cost selection criterion.

1        25. The computer readable medium of claim 16 wherein the  
2 cost associated with the corresponding candidate match for the  
3 prefix pattern, and the cost associated with the candidate  
4 match for the suffix pattern are more heavily weighted than  
5 the cost associated with the candidate match for the value  
6 pattern.

1        26. The computer readable medium of claim 16 wherein the  
2 cost associated with each of the candidate matches for the  
3 first pattern is determined by adding the cost associated with  
4 the corresponding candidate match for the prefix pattern, the  
5 cost associated with the candidate match for the suffix  
6 pattern and the cost associated with the candidate match for  
7 the value pattern.

1        27. The computer readable medium of claim 16 wherein  
2 identifying each candidate match for the first pattern  
3 comprises identifying the candidate match for the prefix

4 pattern, wherein the candidate match for the prefix pattern  
5 defines a first end of a value window, then identifying a  
6 corresponding candidate match for the suffix pattern, wherein  
7 the candidate match for the suffix pattern defines a  
8 corresponding second end of the value window, wherein the  
9 candidate match for the value pattern comprises the symbols  
10 within the value window.

1       28. The computer readable medium of claim 16 further  
2 comprising filtering the candidate match for the value pattern  
3 using a keyword.

4       29. The computer readable medium of claim 16 further  
5 comprising filtering the candidate match for the value pattern  
6 using a regular expression.

7       30. The computer readable medium of claim 16 wherein  
8 identifying candidate matches for the prefix pattern comprises  
9 constructing an edit distance matrix for the prefix pattern  
10 and identifying one or more candidate matches for the prefix  
1 pattern, constructing an edit distance matrix for the suffix  
2 pattern and identifying one or more candidate matches for the  
3 suffix pattern, and identifying a candidate match for the  
4 value pattern between each pair of candidate prefix matches  
5 and candidate suffix matches.